■7群（コンピュータソフトウェア）－ 6編（情報検索とデータマイニング）

# 5章 データ分析の基礎技術

【本章の構成】

本章では以下について解説する.

5-1 分類規則の学習と予測・回帰分析

5-2 因果関係の分析

5-3 クラスタリング

■7 群－6 編－5 章

# 5-1 分類規則の学習と予測・回帰分析

（※準備中）

■7 群－6 編－5 章

# 5-2 因果関係の分析

（執筆者：植野真臣）［2008 年 12 月 受領］

データベースの大量データの因果関係の分析は，一般に「因果モデルの学習」[1] と呼ばれ，(1) 条件付き独立判定による学習と (2) スコアリング法による学習に大別できる．(1) は，特定のモデルを仮定せずに無方向の連続型変数を持つ因果構造の学習にも用いることができるが，(2) は，母数モデルを基調とし，ベイジアンネットワークをデータ発生モデルとして仮定している．

本節では，データベースの解析でよく行われる離散変数データについての因果関係の分析について，(1) 条件付き独立判定による学習，(2) スコアリング法による学習，を説明する．

## 5-2-1 条件付き独立判定による学習

### (1) MWST 法

Chow and Liu (1968) [2] は，離散同時確率分布を相互情報量を用いた木構造で近似する手法を提案しており，MWST（Maximuim Weight Spanning Tree）法と呼ばれる．いま，$N$ 個の離散変数集合 $x=\{x_1,x_2,\cdots,x_N\}$ について，MWST 法は以下のとおりである．

(1) 与えられたデータより，$N$ ($N$－1) /2 個のすべての 2 つの変数ペア（枝と呼ぶ）について，相互情報量

$$I(x_i,x_j)=\sum_{x_i,x_j} p(x_i,x_j)\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)}$$

を求める．

(2) 最も大きな相互情報量の値を示す 2 つの枝を取り出し，木を構成する枝とする．

(3) ループができないならば，次に大きな枝を木に加え，ループができるのであればその枝を捨てる．

(4) ステップ 3 を $N$－1 個の枝が選ばれるまで続ける．

このアルゴリズムの利点として，①二次統計量までしか用いないために，データからの演算が容易で信頼できる，②計算量が高々 $O(n^2)$ のオーダーである，が挙げられる．MWST 法は一変数が単一の親変数を持つ構造を生成するが，複数の親変数を生成するアルゴリズムに Rebane らによって拡張されている [3]．

また，ベイジアンネットワーク分類機の最近の研究では，他の比較的最近提案された学習手法より，MWST 法を改良した手法が高速に高精度の推論ができることが知られている [4]．

### (2) PC アルゴリズム

MWST 法は複雑な因果モデルを単純な木構造に縮約し近似する手法で，変数間の因果関係の正確な分析は難しい．Sprites らにより，より正確な因果関係の分析手法として，以下のような PC 探索アルゴリズムが提案されている [5]．

(1) $N$ 個の変数間にすべて枝を引き，完全無向グラフを構成する．

(2) すべての 2 つの変数ペアについて他の変数を所与として条件付き独立テストを行い，周辺独立もしくは条件付き独立の場合に枝を消去する（ただし，所与とする親変数集合の数を 0 から一つずつ増やしながら条件付き独立テストを行うことにより計算量を減らし

ている).

(3) $(X, Y)$ と$(Y, Z)$ が隣接するが $(X, Z)$ が隣接しない 3 変数 $(X, Y, Z)$ について，$Y$ を所与として $X$ と $Z$ が従属であれば，$X \rightarrow Y \leftarrow Z$ と同定する．$F'$ を部分的に方向づけられたグラフとする.

(4) 以下を $F'$ の枝がこれ以上方向づけられなくなるまで繰り返す.

a：$F'$ において，$X \rightarrow Y$ が存在し，$Y$—$Z$ が存在し，$X$ と $Z$ が隣接していないとき，$Y \rightarrow Z$ である.

b：$X$ から $Y$ へのパスが存在し，$X$—$Y$ が $F'$ にあるとき，$X \rightarrow Y$ とする.

c：4 つの変数 $X, Y, Z, W$ において $Y \rightarrow Z$，$W \rightarrow Z$ があり，$X$—$Y$，$X$—$Z$，$X$—$W$ があるとき，$X \rightarrow Z$ とする.

最悪の場合，ステップ(2)と(3)での計算量が変数の数 $N$ に対して指数爆発する場合もありうることが難点であるが，正確な構造が得られる保証はある．また，完全な有向グラフを得られないかもしれないが，理論的に整合性のある有向枝を部分的に構成することができ，有効である．また，2 つの変数が共通の潜在変数や変数選択バイアスを考慮したグラフ表現 PAG（Partial Ancestral Graphs）を構築するための Fast Causal Inference（FCI）アルゴリズムが提案され，現在も研究が続けられている [5].

### (3) 逐次消去アルゴリズム

PC アルゴリズムでは，真の構造で変数間に枝が多く付く場合には，計算量が指数的に増大してしまうことが欠点であった．植野は，変数間のすべての枝が張られた完全グラフから，一つの枝を消去したときとそうでないときのエントロピーを検定することによる逐次消去アルゴリズム：SAE（Sequential Arcs Elimination）algorithm を提案している [6]．これにより，計算量をデータ数の定数倍オーダーに抑えることができることが特徴で，最近でも，親変数が非常に多い変数を持つ構造の推定に非常に有効であることなどが報告されている.

## 5-2-2 スコアリング法による学習

### (1) Dirichlet Prior Score Metric (DPSM)

スコアリング法による学習では，確率構造にベイジアンネットワークモデルを仮定し，ベイジアンネットワークの条件付き確率をパラメータとしてスコアリングを導出する．最も有名なスコアリング法は，条件付き確率パラメータ $\Theta_{Bs}$ の事前分布に共役事前分布であるディレクレ分布を仮定してデータ $X$ を所与としたときの構造 $B_S$ の予測分布を導出した以下の Dirichlet Prior Score Metric（DPSM）である [7].

$$p(\mathbf{X}|B_S) = \int_{\Theta Bs} p\left(\mathbf{X}\middle|\Theta_{Bs}, B_S\right) p\left(\Theta_{Bs}\right) d\Theta_{Bs} \qquad (2 \cdot 1)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma\left(n'_{ij}\right)}{\Gamma\left(n'_{ij} + n_{ij}\right)} \prod_{k=0}^{r_i - 1} \frac{\Gamma\left(n'_{ijk} + n_{ijk}\right)}{\Gamma\left(n'_{ijk}\right)}$$

ここで，$n_{ijk}$ は親ノードが $j(= 1, \cdots, q_i)$ 番目のパターンをとったときの変数 $i(= 1, \cdots, N)$ が値 $k(= 0, \cdots, r_i - 1)$ をとったサンプル数を示し，$n'_{ijk}$ は $n_{ijk}$ に対応するハイパーパラメータを示す．DPSM を最大にするベイジアンネットワークの確率構造を探索すればよい．特に，DPSM で $n'_{ijk} = 1,\ (i = 1, \cdots, N,\quad j = 1, \cdots, q_i,\quad k = 0, \cdots, r_i - 1)$（一様分布）を仮定した場合，よく

知られた K2 アルゴリズム [7] に採用されたスコアリングとなる．

### (2) Minimum Description Length (MDL)

ベイジアンネットワークのスコアリングの一つとして，DPSM のハイパーパラメータをすべて $n'_{ijk}1/2,\ (i=1,\cdots,N,\quad j=1,\cdots,q_i,\quad k=0,\cdots,r_i-1)$ と設定して漸近展開を行うことによって，以下の MDL（Minimum Description Length）が導くことができる [8]．

$$I(B_S,\mathbf{X})=\ln p(B_S)+\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\left[n_{ijk}\ln\frac{n_{ijk}}{n_{ij}}\right]-\frac{\sum_{i=1}^{N}q_i(r_i-1)}{2}\ln n \qquad (2\cdot 2)$$

ここで，$n=\sum_{j=1}^{q_i}n_{ij}$ となる．その後，植野 [9] により，$n'_{ijk}=1/2$ の条件がより緩和され，条件 $2\pi(\ln n)^2-n<n'<2\pi\exp(2n)-n$ で式(2･2)に収束することが証明されている．

### (3) BDe (Bayesian Dirichlet Equivalence)

Heckerman らは [10]，K2 に採用された DPSM のハイパーパラメータ $n'_{ijk}=1$ では，Likelihood Equivalence（構造の識別可能性との対応）を持たないことを指摘し，ノードについてのハイパーパラメータの値の和が一定になることが条件であることを示している．更に事前のユーザの知識構造 $B_S h$ を所与としてハイパーパラメータ制約 $n'_{ijk}=n'p(x_i=k,\ \prod_i=j\,|B_S\ h)$（ここで，$n'$ は，equivalent sample size と呼ばれ，仮説である構造 $B_S\,h$ の学習への重みを意味している）を提案し，Likelihood Equivalence を持ち，ユーザの事前知識を反映させることができる DPSM として Bayesian Dirichlet equivalence（BDe）と呼んでいる．また，Buntine [11] が先に提案している一様事前分布のハイパーパラメータ $n'_{ijk}=n'/(r_iq_i)$ を BDe の特殊形として，BDeu（u は uniform の u）と呼んでいる．近年，BDeu への注目は増しており，経験的ベイズ手法を用いて BDeu の equivalent sample size を決定することにより，最も予測効率の高い学習が行えることなどが報告されている [9]．

### (4) NPC (Necessary Path Condition) アルゴリズム

条件付き独立判定とスコアリング法を組み合わせた学習法も提案されており，NPC（Necessary Path Condition）と呼ばれている [12]．

**■参考文献**


1) C. Glymour and G.F. Cooper：Computation, Causation, and Discovery, The AAAI Press, 1999.
2) C.K. Chow and C.N. Liu：“Approximating discrete probability distributions with dependence trees,”IEEE transactions on information theory, vol.14, pp.462-467, 1968.
3) G. Rebane and J. Pearl：“The Recovery of Causal Poly-trees from statistical data,”Proc. Uncertainty in Artificial Intelligence, pp.222-228, 1987.
4) J.Cheng and R.Greiner：“Comparing Bayesian Network Classifers,”Proc. Uncertainty in Artificial Intelligence, pp.101-108, 1999.
5) P. Sprites, C. Glymour and R. Scheines：Causation, Prediction, and Search, New York, Springer-Verlag, 1993.
6) 植野真臣:“意思決定アプローチによる Bayesian Network の因果モデル構築,”人工知能学会論文誌, vol.11, no.5, pp.49-58, 1996.
7) G.F. Cooper and E. Herskovits：“A Bayesian Methods for the induction of probabilistic networks from data,” Machine Learning, vol.9, pp.309-347, 1992.
8) J. Suzuki：“A Construction of Bayesian networks from Databases on an MDL Principle,”Proc. Uncertainty in Artificial Intelligence, pp.266-273, 1993.
9) M.Ueno：“Learning likelihood equivalence Bayesian networks using an empirical Bayesian approach,” Behaviormetrika, vol.35, no.2, pp.115-135, 2008.

10) D. Heckerman, D. Geiger, and D.M.Chickering：“Learning Bayesian networks: The combination of knowledge and statistical data,” Machine Learning, 20(3), pp.197-243, 1995.
11) W. Buntine：“Theory refinement on Bayesian networks,” Proc. Uncertainty in Artificial Intelligence, pp.52-60, 1991.
12) H.Steck：“Constraint-Based Structural Learning in Bayesian Networks using Finite Data Sets,” Institut fur Informatik der Technischen Universitat Munchen, 2001.

■7 群－6 編－5 章

# 5-3 クラスタリング

（執筆者：上田修功）［2009 年 11 月 受領］

クラスタリングとは，あらかじめ定義されたサンプル間の距離（非類似度）に基づき，与えられたサンプル集合をいくつかのグループに分割する手法の総称で，これまで，統計，機械学習，パターン認識，人工知能，データマイニングなどの分野で様々な手法が提案され，応用されている [1].

クラスタリング対象の各サンプルは，通常，ユークリッド空間上の点（実ベクトル）とするが，頻度データやシンボリックなサンプルを対象とするクラスタリング手法も提案されている．例えば，サンプルがグラフのノード集合として与えられ，ノード間のリンクの接続関係に基づくグラフクラスタリング手法がある [2].

サンプル集合 $D=\{x_1,\cdots,x_N\}$ に対し，最も古典的かつ直観的な手法である最短距離に基づく階層的クラスタリングでは，2 つのクラスタ間距離を定義し，1 つのサンプルからなる $N$ クラスタを初期値とし，最もクラスタ間距離が小さい 2 つのクラスタを逐次併合し，階層的なクラスタ構造を求める．

実ベクトルを対象とする分割型のクラスタリング手法の代表手法として，$K$-means 法がある．サンプル集合 $D=\{x_1,\cdots,x_N\}$ に対し，$K$ 個のクラスタ代表点からなる代表点集合 $Y=\{y_1,\cdots,y_K\}$ を導入する．サンプル $x_i$ をユークリッド自乗距離 $d(x_i,y_j)=\|x_i-y_j\|^2$ が最小となるクラスタ $j$ に帰属させるというルールに従い，$D$ の要素を排他的に $K$ 分割する．通常，$K$ は $N$ に比べ十分小さい．$Y$ を次式の目的関数の最小化問題として解く．

$$J(D,Y)=\sum_{i=1}^{N}\sum_{j=1}^{K}d(x_i,x_j) \qquad (3\cdot 1)$$

上記目的関数は，$D$ を距離 $d$ のもとで最良近似する $Y$ を求めていることと解釈できる．つまり，式(3･1)は，$N$ 個のベクトルを $K$ 個のベクトルで近似表現するベクトル量子化の平均量子化誤差を表す．$K$-means 法の具体的手順を以下に示す．

(1) $k$ 個のクラスタの代表点 $y_1,\cdots,y_K$ を初期化する．

(2) $x_i, i=1,\cdots,N$ に対し，$x_i$ を $d(x_i,y_j)$ が最小となる $j$ に割り当てる．

(3) もし，代表点への割当ての変化がなければ終了．さもなくば，各代表点の値をその代表点に割り当てられたサンプルの平均値として更新し，ステップ 2 へ．

上記アルゴリズムは，必ず収束するが，式(3･1)の最小値を与える $Y^*$ が得られる保証はなく，ステップ 1 の代表点の初期値に近い局所最適値に収束する．換言すれば，$K$-means 法の結果は初期値設定に大きく依存する．この初期値依存性を解決するための効果的な手法も提案されている [3].

Kmeans 法では，排他的なクラスタリング，すなわち，各サンプルは必ずいずれか一つのクラスタに決定論的に帰属することになるが，各クラスタへの帰属度を確率値として算出可能なクラスタリング手法もある．以下，これについて概説する．

$j$ クラスタのサンプル $x$ を確率分布 $p_j(x;\ \theta_j)$ からの実現値と仮定する．ここで，$\theta_j$ はクラス

タ $j$ のモデル分布のパラメータを表す．クラスタの混合比を $\alpha_1,\cdots,\alpha_K$ $\left(\alpha_j \geq 0, \sum_j \alpha_j = 1\right)$ とすると，サンプル $x$ の分布は，

$$P(x;\Theta) = \sum_{j=1}^{K} \alpha_j P_j(x;\ \theta_j) \qquad (3・2)$$

と書ける．換言すれは，上記確率モデルは，まず，$(\alpha_1,\cdots,\alpha_K)$ をパラメータとする多項（$K$ 項）分布でクラスタ $j$ を選択し，次いで，そのクラスタの要素分布 $P_j(x;\ \theta_j)$ からサンプル $x$ が生成されたと仮定する．このように，複数の要素分布からなる確率モデルを混合モデルと呼ぶ．混合モデルの構成要素である要素分布は応用に応じて設定される．

未知パラメータ $\Theta = \{\alpha_1,\cdots,\alpha_K,\ \theta_1,\cdots,\theta_K\}$ は，最尤推定法に基づき，制約条件 $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$ のもとで，次式の対数尤度関数の最大化問題として求められる．

$$\mathcal{L}(\Theta;D) = \log\prod_{i=1}^{N} P(x_i;\Theta) = \sum_{i=1}^{N} \log\sum_{j=1}^{K} \alpha_j P_j(x;\ \theta_j) \qquad (3・3)$$

上記目的関数はパラメータ $\Theta$ に関し非線形ゆえに，閉じた形で解くことはできず，ニュートン法などの非線形数値解法で逐次的に解くことになるが，混合モデルに対するより簡便なパラメータ推定法として EM アルゴリズム [4] が著名である．

パラメータ推定値を $\Theta^*$ とすると，サンプル $x$ のクラスタ $j$ への帰属事後確率は，ベイズルールより $P(j|x) = \alpha_j^* P_j\left(x;\Theta^*\right)/P\left(x;\Theta^*\right)$ と計算でき，各クラスタへの帰属度が確率的に算出できる．排他的なクラスタリングを必要とする場合は，$x$ を事後確率が最大の $j$ に割り当てればよい．要素分布として多項分布モデルを仮定すると，文書データの単語出現データのような頻度データのクラスタリングに応用でき，実際，データマイニングの分野で，文書をスポーツ，音楽，政治などのトピックにクラスタリングする手法としてこの混合モデルアプローチが用いられている．また，混合モデルアプローチにおける混合数 $K$ はあらかじめ設定されるが，近年，ノンパラメトリックベイズ法の枠組みで，クラスタリング過程で自動的に適切な $K$ を求めることも可能である．

**■参考文献**


1) A.K. Jain, M.N. Murty, and P.J. Flynn：“Data clustering,” A Review, ACM Computing Surveys, vol.31, no.3, 1999.
2) C.H.Q. Ding, X. He, H. Zha, M. Gu, and H.D. Simon：“A min-max cut algorithm for graph partitioning and data clustering,” In Proc. of the IEEE Int'l Conf. on Data Mining, pp.107-114, 2001.
3) N. Ueda and R. Nakano：“A new competitive learning approach based on an equidistortion principle for desining optimal vector quantizers,” Neural Networks, vol.17, no.8, pp.1211-1227, 1994.
4) A.P. Dempster, N.M. Laird, and D.B. Rubin：“Maximum likelihood for incomplete data via the EM algorithm,” Journal of the Royal Statistics Society (B), vol.39, no.1, pp.1-38, 1977.